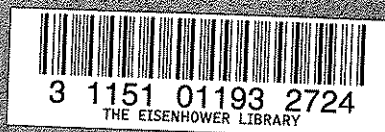


HA
29
.M 31
1953



LSC

9349.11

R. M. McClung

First aid for pet projects injured in the
lab or on the range or what to do until the
statistician comes.

China Lake

1953

UNCLASSIFIED

NOTS TECH. MEMO. No. 1113

FIRST AID FOR PET PROJECTS
INJURED IN THE LAB OR ON THE RANGE
OR
WHAT TO DO UNTIL THE STATISTICIAN COMES

By
R. M. McClung
Aviation Ordnance Department

DATE WRITTEN: 29 January 1952

DATE REPRINTED: 21 January 1953

U. S. NAVAL ORDNANCE TEST STATION, INYOKERN
China Lake, California

UNCLASSIFIED

FOREWORD

This memorandum is transmitted for information purposes only. It does not represent the official views or final judgment of the Naval Ordnance Test Station, and the Station assumes no responsibility for action taken on the basis of its contents.

Wm B. McLean

ABSTRACT

The concept of confidence, the use and computation of confidence and tolerance limits, and a method for determining the significance of the difference in two or more sample means are discussed in an elementary manner.

TABLE OF CONTENTS

	Page
Acknowledgement	iii
Foreword	v
Abstract	vi
List of Tables	viii
Introduction	1
PART I - The Concept of Confidence	3
PART II - Confidence Limits for a Binomial Distribution	6
PART III - Standard Deviation, Normal Distribution, and Tolerance Limits	23
PART IV - Confidence Limits for an Average	33
PART V - The Significance of an Observed Difference in Averages for Two Series of Tests	37
PART VI - Confidence Limits for a Standard Deviation	49

LIST OF TABLES

			Page
TABLE	I	Confidence Limits (percent) for Binomial Distribution ...	13
TABLE	II	Confidence Limits (percent) for Binomial Distribution ...	14-15
TABLE	III	Confidence Limits (percent) for Binomial Distribution ...	16-17
TABLE	IV	Confidence Limits (percent) for Binomial Distribution ...	18-19
TABLE	V	Upper or Maximum Confidence Limit (percent) For a Series of Tests Without a Failure ...	20
TABLE	VI	Number of Tests to be Performed Without a Failure ...	22
GRAPH		Normal Frequency Curve ...	27
TABLE	A	Tolerance Factors for 90 Percent Confidence ...	30
TABLE	B	Tolerance Factors for 95 Percent Confidence ...	31
TABLE	C	Tolerance Factors for 99 Percent Confidence ...	32
TABLE		Student's t Distribution ...	36
GRAPH		Student's t Distribution ...	47
TABLE	B	Factors ...	51

INTRODUCTION

"The practice sometimes followed of consulting the statistician only after the experiment is completed and asking him 'what he can make of the results' cannot be too strongly condemned. It is essential to have the experiment in a form suitable for analysis and in general this can be attained by designing the experiment in consultation with the statistician, or with due regard to the statistical principles involved."--R. C. Bowden, Director of Ordnance Factories (Explosives), Ministry of Supply, Great Britain.

An electronic experimenter possesses a sine wave generator which can economically manufacture millions or even billions of essentially identical repetitive input signals for a circuit that he may be developing. He can therefore base his conclusions relative to the performance of this circuit with respect to input signal and with respect to changes in circuit components upon a sufficient number of observations so that the conclusions can be accepted without reservation.

The developer of rocket ordnance material is not so fortunate. The cost in time and money of a guided missile or even a small component such as a fuze is so great that conclusions which may involve the expenditure of millions of dollars and the strengthening or weakening of the national security must, of necessity, be made upon the basis of very small numbers of tests. Great strides have been made in the last few years in the science of statistics; many of these advances have been made since the majority of practicing scientists and engineers received their formal professional training. Although it may be tacitly assumed that professional personnel engage in outside reading which advises them of all significant advances in all professional fields which may conceivably relate to their own, the writer questions the validity of this assumption, and has noted that professional personnel generally maintain their proficiency in their own specialty only and may very likely forget even the basic essentials in the other professions which were taught in their undergraduate orientation courses.

The advice and assistance of a professional statistician are the first prerequisites to the performance of a successful series of tests of a small number of expensive items from which important

conclusions must be drawn. It is the engineer, however, who must determine when to consult the statistician, just as it is the patient who must determine when to visit the physician. As assistant for engineering operations in the Rocket Division, the writer noted that many of the engineers in that and other divisions were not thinking or planning statistically; the statistician was a gentleman who "worked upstairs somewhere" with a lot of tables which could be used to add lustre to the phraseology of a final report.

The writer concluded that the valuable assistance of the professional statistician would not be utilized until each engineer concerned with the planning or the execution of ordnance experimental programs incorporated a statistical attitude into his daily thinking. In an effort to facilitate this incorporation, the writer conducted a series of statistical discussions for project engineers and prepared the "first aid notes" therefor. If the title leads any reader to believe that a statistician should be consulted only after difficulties have been encountered in the experimental program, this is regrettable, but the writer suspects that this reader would not have consulted the statistician at all otherwise so the net effect may be beneficial, at least with respect to future programs.

The style of this memorandum is deliberately informal and pedagogical. No attempt has been made to provide a comprehensive introduction to the field of statistics; conventional first aid notes are not an introduction to the field of medicine. This is a sample, in the culinary sense, which it is hoped will whet the appetite of the reader.

PART I

"We see that the theory of probabilities is at bottom only commonsense reduced to calculation: it makes us appreciate with exactitude what reasonable minds feel by a sort of instinct."--P. S. Laplace

Purpose:

To discuss the concept of confidence.

a. The owl has gained a reputation for wisdom, which ornithologists say is undeserved, by maintaining an astute look and never making an incorrect statement. It should be noted, however, that he never makes ANY statements. The scientific experimenter finds that an astute look is not a liability, but his supervisor requires him to make statements and draw generalized conclusions on the basis of his experiments.

b. Obviously, the experimenter who makes definitive statements relative to the performance that may be expected of a large number of rockets, solely on the basis of the test results accumulated from five or ten firings, will sometimes be wrong in these statements. A confidence level or confidence coefficient is a state of mind and reflects how anxious the owner of that mind is to avoid being wrong in his statements or conclusions.

c. A cautious, conservative person who buys safe investments, wears a belt and suspenders, and qualifies his statements carefully is operating on a high confidence level. He is certain he won't be wrong very often. If he is wrong once in 100 times, he is operating on a 99 percent confidence level. A less conservative person who takes more chances will be wrong oftener and hence he operates on a lower confidence level. If he is wrong one time in 20 statements, he is operating on a 95 percent confidence level. The confidence level, therefore, merely specifies the percentage of the statements that a person expects to be correct. If the experimenter selects too high a confidence level, his test program becomes prohibitively expensive in time and money before he makes any very precise conclusions. If the confidence level is too low, precise conclusions are easily reached but these

conclusions will be wrong too frequently, and this in turn is too expensive in time and money if a large quantity of the item is made on the basis of erroneous conclusions. There is no ready answer to this dilemma. I favor using a low confidence level when tests are exploratory and progressively raising the confidence level as the finished item is evaluated. This practice is not always applicable, however.

d. Tables will be presented in the other parts of this memorandum which permit the experimenter to determine what conclusion he can draw for the confidence level that he selects. If he selects a confidence level of 95 percent, he uses the 95 percent confidence table and the conclusions that he draws in accordance with this table will be in error only one time in twenty, on the average.

e. Many statisticians prefer the term Confidence Coefficient to Confidence Level; the meaning is the same.

f. Some authors prefer to consider the confidence level as representing the probability that a given statement is correct. The reader may adopt this concept if he wishes, but I would not recommend it as I believe that it involves a philosophical concept which eventually hinders an understanding of the subject. A further discussion of this difference in viewpoints is given in a footnote which may be of interest to some readers.

FOOTNOTE

If a coin is flipped the chances are approximately one out of two before the coin lands that it will be heads. After it lands, but even before I look at it, the uncertainty has ended and it is either heads or tails, and it does not seem correct for me to then state that the probability of it being heads is one out of two. Everyone else in the room may have already looked at it and knows that it is tails. If I shuffle a deck of cards and then ask what the chances are that the top card is the ace of spades, the answer is not one in 52; to my way of thinking the question is meaningless. The top card is either the ace of spades or it isn't and no chances are involved. If I state that it is NOT the ace of spades, I know that I am operating on a confidence level of 51/52 or approximately 98 percent; that is, I will be right in 51 out of every 52 such statements

that I make, on the average.

Another individual might prefer to state that there is a probability of 51 in 52 that the top card is not the ace of spades and a statement to the effect that the top card is not the ace of spades is therefore made with a probability of 51/52 or approximately 98 percent of being correct.

In the case of an experimental rocket that may be tested for burning distance, I would prefer to state that the true average burning distance is established by the rocket design and exists whether any rockets are ever built or fired, although no one knows what this distance is. If I do test a few, I can then make a statement relative to this distance, with appropriate limits, and I will be right or wrong, although I do not know if I'm right or wrong. I do know that I will be right in such statements a certain known percentage of the time, say 95 percent, which represents my confidence level, because I got my statement from a 95 percent confidence table. I do not say that my probability of being right on a given statement is 95 percent, as I am either right or wrong, and I prefer not to imply that my ignorance of the true fact creates a probability. Some authors on the subject do refer to the probability of being right on a given statement, however, and for them the confidence level might be defined as the probability of being right in any given statement. In this memorandum, the confidence level will refer to the percentage of his statements that the experimenter may expect to be correct.

PART II

"Mathematics may be compared to a mill of exquisite workmanship which grinds you stuff of any degree of fineness, but nevertheless what you get out depends on what you put in--and as the grandest mill in the world will not extract wheat flour from peas pods so pages of formulae will not get a definite result out of loose data."--T. H. Huxley

Purpose:

To define confidence limits.

a. Confidence levels and confidence limits have been confused in the minds of many professional personnel. A level and a limit are not the same. Confidence limits are merely the computed upper and lower limits of the desired value of a physical quantity. For example, the average height of the Joshua tree might be of interest.

b. I could measure the height of a number of Joshua trees picked at random in a number of localities and compute very accurately the average height of those I had measured. Suppose this value were 15 feet 6 inches and I had measured 20 trees. I might decide to state in a published report that the average height of Joshua trees in general was between 14 and 17 feet, which would be my limits. If someone else said 15 and 16 feet, their limits would be 15 and 16 feet. Obviously the latter limits are rather close for a measurement based upon only 20 trees and these limits would be made by a person operating on a lower confidence level than the person who selected the limits of 14 and 17 feet.

c. In other words, the person who sets his limits quite closely to his observed value will have a lower percentage of his statements correct than a person who is not so precise. If the above limits were obtained from suitable statistical tables, they would be called confidence limits on the average height of the Joshua tree. The difference between the upper and lower limit is known as the confidence interval. If a botanical experimenter had determined that his reputation would not be seriously affected if he were right in only 95 percent of his statements, he would have gotten his limits by means of a 95 percent confidence table and he would refer to the

limits so derived as 95 percent confidence limits and the interval between as a 95 percent confidence interval. He therefore states positively that the average height of the Joshua tree is between these limits, but he knows that he will be wrong in 5 percent of such statements.

d. Since most tables yield central confidence intervals, he also knows that he will be wrong in 2 1/2 percent of his statements because the true value was equal to or less than his lower confidence limit and wrong in 2 1/2 percent of his statements because the true value was equal to or greater than his upper confidence limit.

e. It is therefore obvious that he would be operating on a 97 1/2 percent confidence level if he had made a statement utilizing only one of the limits. Such a statement might be: "The average height of the Joshua tree is less than _____ feet." Even though he used the same numerical value as before, he knows that 97 1/2 percent of such statements will be correct, as the true value will be equal to or greater than the limit for only 2 1/2 percent of his statements. If he had wished to stay on the 95 percent confidence level, and specify only one limit, the numerical value of the maximum limit could be a little lower because he can afford to be wrong in 5 percent of his statements because the true value was equal to or greater than his limit. It is generally better to retain the same confidence level and compute the desired limit or pair of limits accordingly.

f. In the case of the Joshua tree, the more conventional approach described in paragraph c will yield more informative statements for any specified confidence level than the use of a single confidence limit as illustrated in paragraph e. As will be shown later, however, for certain problems in ordnance engineering, the specification of only one limit may be desirable. Inasmuch as the words "upper" and "lower" as applied to confidence limits have long been associated with the specification of a confidence interval to which the confidence level applied, I wish to propose the terms "maximum confidence limit" and "minimum confidence limit" to define limits which will be used in a singular sense with no reference to each other or to a confidence interval. (The maximum confidence limit and the upper confidence limit are therefore not synonymous for a specified confidence level.)

g. The technique for computing confidence limits for an average from observed data is given in Part IV of this memorandum, as some terms must be defined in Part III before this operation can be explained.

h. For a little practice in the use of confidence limits, consider the type of tests in which only two answers are possible. The drop testing of light bulbs might be a case: either the bulb breaks or it doesn't. The gauging of metal parts is another; the firing of detonators on a specified electrical current is a third. Either an anticipated event occurs or it doesn't; no measured data are involved, the answer is yes or no. Assume that 10 point detonating bomb fuzes are dropped and three fail to function. The best estimate of the failure rate of an infinite number of identical fuzes is the one observed; that is, 30 per cent, but what may be said relative to a confidence interval for the failure rate?

i. Referring to Table III, it may be seen that the 95 per cent confidence interval is from 7 per cent to 65 per cent, the lower and upper confidence limits respectively. A statement suitable for a report might be, "A sample failure rate for the subject fuze of 30 per cent was observed for ten samples, and the writer is 95 per cent confident that the true failure rate lies between 7 per cent and 65 per cent." Ninety-five of every one hundred such statements will be correct on the average. The specification of the number of fuzes tested is not essential but it takes little space and enables a reader to check the statement. Unless the experimenter knows the use that is to be made of his report, this statement employing a confidence interval is the most informative and should be given.

j. If the tests in the preceding paragraph were made upon ten fuzes picked at random from a lot that had been in storage, for the purpose of deciding whether to issue or rework the fuzes, the man responsible for making the decision might be more interested in the following singular limit statement which could be made after reference to Table II: "A sample failure rate for the subject fuze of 30 per cent was observed for ten samples, and the writer is 95 per cent confident that the true failure rate is less than 61 per cent." (Since I have invented the terms "maximum" and "minimum" confidence limits, their use is not recommended in a formal report, but the recommended statement permits no ambiguity. Sixty-one per cent is the maximum confidence limit.)

k. Conversely, if the fuzes were models of an experimental fuze, the man responsible for deciding whether to make more fuzes to this design might be more interested in a singular limit type of statement as follows: "A sample failure rate for the subject fuze of 30 per cent was observed for ten samples, and the writer

is 95 per cent confident that the true failure rate will exceed 9 per cent if a large number of fuzes are made in accordance with this design." This statement is also taken from Table II.

l. In either case, the use of a singular confidence limit permitted the statement to be a little more precise in the field of interest. In other examples that might be cited, the gain in precision is more appreciable. It must be EMPHASIZED, however, that the experimenter must select his confidence level and must decide whether he is interested in a confidence interval or a singular limit, and in this case which limit, before he analyzes his data. If he favors one technique or the other after he sees the data, he is biasing the probability involved in the computation of the tables and he will lower his over-all batting average below the indicated confidence level as these tables are computed upon the assumption that they will be used in a consistent manner. Suppose, for example, that he observes no failures in ten tests. If he had intended to determine a maximum confidence limit only, he may say with 95 per cent confidence (Table II) that the true failure rate is less than 26 per cent. If he had intended to specify a confidence interval, he must say that the true failure rate lies in the interval of 0 per cent to 31 per cent. He cannot jump from an interval type of analysis to a singular limit type of analysis only when he observes no failures.

m. Personally I favor the use of a singular limit technique for ordnance component testing as the gain in precision is considerable when no failures are observed, and this gain in precision is legitimate if this analysis is also used when failures are observed. When failures are observed, there is a loss in precision if the singular limit technique is employed but this is probably of no consequence as the experimenter is only interested in how high (or how low) the true failure rate might be or he would not have adopted the single limit type of analysis in the first place.

n. Tables I and IV are used in a similar manner for other confidence levels that the experimenter prefers, although it should be noted that the 95 per cent confidence level is quite popular with statisticians and experimenters.

o. The ordnance experimenter who uses Tables I through IV soon discovers that he does not often have his data for series of tests consisting of 10, 20, or 50 samples. He may have had sufficient funds to test only 6 units, for example. He has observed

no failures and is exhibiting a pardonable pride in this achievement. But what definitive statement can he make in his report? With what assurance can he recommend comparatively large scale manufacture of the item?

p. I have prepared Table V for the purpose of answering these questions. If the experimenter believes that a 95 per cent confidence level is desirable and if he has decided before performing the test that he is going to make a singular limit statement for the maximum limit, he makes the statement that the true failure rate will be less than 39 per cent for a very large number of items identical to those he has tested. He makes this rather unencouraging statement with the knowledge that he will be wrong in 5 per cent of such statements! Even if he were willing to operate on an 80 per cent confidence level (and this is not recommended), he can only state that the true failure rate is less than 24 per cent.¹ It is small wonder that many projects which appear to offer great promise on the basis of a few exploratory tests later lead to disappointment. Conversely, it should be remembered that the best estimate of the true failure rate is the one observed, 0 per cent in this example. If the experimenter has faith, which is not amenable to statistical analysis, in his item and his own ability to remedy such deficiencies as may later become apparent, he may proceed with enthusiasm and high emotional, if not statistical, confidence.

q. Realizing that too few tests in a series results in a very broad confidence interval or a high maximum confidence limit, the experimenter may ask how many successful tests without a failure are necessary before he can claim to have met a maximum confidence limit specified by the Bureau or his supervisor. I have prepared Table VI for this purpose. It may be used for the estimating of quantities of items required for test purposes.

r. The experimenter should examine very critically any requirements stated in statistical terms that may be given him by higher authority. For example, the requirement of "99 per cent functioning at 95 per cent confidence" reads very nicely but it might be prohibitively costly to demonstrate for anything more expensive than machine gun bullets. Since the higher authority is interested in only how high the failure rate is, a single confidence limit approach is legitimate. The experimenter will have to test

¹This value was computed from formula at bottom of page 21.

299 items without a failure (Table VI) in order to demonstrate that he has met the requirement. This is no mean feat and will not often be achieved unless the true failure rate is somewhat under 1 per cent.

s. If the experimenter observes 10 failures in 1000 tests, he has observed a functioning rate of 99 per cent, it is true, but he cannot say that he has met the requirement at the required confidence level. He can say that the true failure rate is less than 2 per cent, yet the chances are that the originator of the requirement would be delighted if he knew that even 100 items had been tested with only a 1 per cent observed failure rate, although the 95 per cent confidence statement would then be that the true failure rate was less than 4 per cent.

t. The originator of the requirement in paragraph r probably hoped that the failure rate of the items tested would be less than one per cent. This is a legitimate hope and should be stated in this manner. To restate the hope as a requirement in statistical terms automatically shifts the emphasis from the failure rate of the samples to be tested to the expected maximum failure rate of an infinite number of items and may pose an insurmountable and probably unintended problem for the experimenter. The experimenter should immediately object to such a requirement and propose an alternate specification in precise statistical terms which he believes can be achieved and state his reasons therefor. He will thereby retain the respect of his superiors whereas he may only incur censure and criticism if he accepts the project without protest and then attempts to justify the performance of his item and lower the required specifications in his final report.

u. The experimenter should not overlook the possibility of utilizing his test data to establish more than one conclusion. For example, a certain maximum confidence limit may have been specified with respect to a rocket motor exploding upon ignition before the motor in question may be used on an aircraft. The cost of the number of motors that must be fired to demonstrate this limit is quite high, but there is no reason why these motors may not be the same ones used to determine the ground launched dispersion and other performance characteristics of the rocket which must be measured in an over-all evaluation program. The motor explosion is sufficiently spectacular to warrant attention by someone whenever it occurs, so no particular preparation for observing this phenomenon is needed. In other cases, however, the inclusion of an extra camera or other instrument may permit the accumulation of data for a considerable number of items that are being tested primarily for other purposes.

The decrease in the confidence interval that occurs when the number of items tested is increased is spectacular for small numbers of items and warrants the extra planning involved.

v. The tables given in this part may also be used to draw conclusions as to the serviceability of ammunition which has been in storage for a long period of time or under adverse conditions. For the results to be valid, it is essential that the selection of the items to be tested be made completely at random.

CONFIDENCE LIMITS (PER CENT) FOR BINOMIAL DISTRIBUTION--TABLE I

Confidence Level--80 per cent for an interval, i.e., a pair of limits.

--90 per cent for a single limit, either max. or min.

(Select level and type of limit desired before analyzing data.)

PER CENT FAILURES IN SAMPLE	SIZE OF SAMPLE			
	5	10	25	50
0	0 37	0 21	0 9	0 5
2				1 6
4			1 13	1 10
6				3 13
8			2 20	4 16
10		1 34		5 18
12			4 25	6 20
14				8 22
16			7 29	9 24
18				11 27
20	3 59	5 45	10 34	13 29
22				14 31
24			13 38	16 33
26				18 35
28			16 43	19 37
30		11 55		21 39
32			19 47	23 41
34				25 44
36			23 51	27 46
38				29 48
40	11 75	18 65	26 55	30 50
42				32 52
44			30 59	34 54
46				36 56
48			33 63	38 58
50*		27 73		40 60

*If percentage observed in sample exceeds 50 per cent, read 100 minus the percentage observed and subtract each confidence limit from 100.

CONFIDENCE LIMITS (PER CENT) FOR BINOMIAL DISTRIBUTION--TABLE II

Confidence Level--90 per cent for an interval, i.e., a pair of limits.

--95 per cent for a single limit, either max. or min.

(Select level and type of limit desired before analyzing data.)

PER CENT FAILURES IN SAMPLE	SIZE OF SAMPLE					
	5	10	25	50	100	1000
1	0 45	0 26	0 11	0 6	0 3	0 1
2				0 8	0 4	0 2
3					1 5	1 3
4			0 16	1 11	1 7	2 4
5					2 8	3 5
6					3 9	4 6
7				2 14	3 11	5 7
8			2 23	3 17	4 12	6 9
9					5 14	7 10
10		0 39		4 19	5 15	8 11
11					6 16	9 12
12			3 28	5 22	7 17	10 13
13					8 18	11 14
14				7 24	9 19	12 15
15					9 21	13 16
16			6 33	8 27	10 22	13 17
17					11 23	14 18
18				9 29	12 24	15 19
19					13 25	16 20
20	1 66	4 51	8 37	11 31	13 26	17 21
21					14 27	18 22
22				13 33	15 28	19 23
23					16 29	20 24
24			11 42	14 35	17 31	21 25
25					18 32	22 26
26				16 38	19 33	23 27
27					20 34	24 28
28			14 46	18 40	21 35	25 29
29					22 36	26 30
30		9 61		19 42	23 37	27 31
31					24 38	28 33
32			17 51	21 44	25 39	29 34
33					26 40	30 35
34				23 46	27 41	31 36
35					28 42	32 37
36			20 55	25 48	28 43	33 38
37					29 44	34 39
38				27 50	29 45	35 40
39					30 46	36 41
40	8 81	15 70	23 59	28 52	31 47	37 42
					32 48	38 43

CONFIDENCE LIMITS--TABLE II (Continued)

PER CENT FAILURES IN SAMPLE	SIZE OF SAMPLE					
	5	10	25	50	100	1000
41					33 49	39 44
42				30 54	34 50	40 45
43					35 51	41 46
44			27 62	32 56	36 52	42 47
45					37 53	43 48
46				34 58	38 54	44 49
47					39 55	45 50
48			30 66	36 60	40 56	46 51
49					41 57	47 52
50*		22 78		38 62	42 58	48 52

*If percentage observed exceeds 50 per cent, read 100 minus percentage observed, and subtract each confidence limit from 100.

TECHNICAL MEMORANDUM No. 1113

CONFIDENCE LIMITS (PER CENT) FOR BINOMIAL DISTRIBUTION--TABLE III
 Confidence Level--95 per cent for an interval, i.e., a pair of limits.
 --97 1/2 per cent for a single limit, either max.
 or min.

(Select level and type of limit desired before analyzing data.)

PER CENT FAILURES IN SAMPLE	SIZE OF SAMPLE				
	10	20	50	100	1000
0	0 31	0 17	0 7	0 4	0 0
1				0 5	0 2
2			0 11	0 7	1 3
3				1 8	2 4
4			0 14	1 10	3 5
5		0 25		2 11	4 7
6			1 17	2 12	5 8
7				3 14	6 9
8			2 19	4 15	6 10
9				4 16	7 11
10	0 45	1 31	3 22	5 18	8 12
11				5 19	9 13
12			5 24	6 20	10 14
13				7 21	11 15
14			6 27	8 22	12 16
15		3 38		9 24	13 17
16			7 29	9 25	14 18
17				10 26	15 19
18			9 31	11 27	16 21
19				12 28	17 22
20	3 56	6 44	10 34	13 29	18 23
21				14 30	19 24
22			12 36	14 31	19 25
23				15 32	20 26
24			13 38	16 33	21 27
25		9 49		17 35	22 28
26			15 41	18 36	23 29
27				19 37	24 30
28			16 43	19 38	25 31
29				20 39	26 32
30	7 65	12 54	18 44	21 40	27 33
31				22 41	28 34
32			20 46	23 42	29 35
33				24 43	30 36
34			21 48	25 44	31 37
35		15 59		26 45	32 38

CONFIDENCE LIMITS (PER CENT) FOR BINOMIAL DISTRIBUTION--TABLE III
(Continued)

PER CENT FAILURES IN SAMPLE	SIZE OF SAMPLE				
	10	20	50	100	1000
36			23 50	27 46	33 39
37				28 47	34 40
38			25 53	28 48	35 41
39				29 49	36 42
40	12 74	19 64	27 55	30 50	37 43
41				31 51	38 44
42			28 57	32 52	39 45
43				33 53	40 46
44			30 59	34 54	41 47
45		23 68		35 55	42 48
46			32 61	36 56	43 49
47				37 57	44 50
48			34 63	38 58	45 51
49				39 59	46 52
50*	19 81	27 73	36 64	40 60	47 53

* If percentage observed in sample exceeds 50 per cent, read 100 minus percentage observed and subtract each confidence limit from 100.

CONFIDENCE LIMITS (PER CENT) FOR BINOMIAL DISTRIBUTION--TABLE IV
Confidence Level--99 per cent for an interval, i.e., a pair of limits

--99 1/2 per cent for a single limit, either max. or min.

(Select level and type of limit desired before analyzing data.)

PER CENT FAILURES IN SAMPLE	SIZE OF SAMPLE				
	10	20	50	100	1000
0	0 41	0 23	0 10	0 5	0 1
1				0 7	0 2
2			0 14	0 9	1 3
3				0 10	2 4
4			0 17	1 12	3 6
5		0 32		1 13	3 7
6			1 20	2 14	4 8
7				2 16	5 9
8			1 23	3 17	6 10
9				3 18	7 12
10	0 54	1 39	2 26	4 19	8 13
11				4 20	9 14
12			3 29	5 21	9 15
13				6 23	10 16
14			4 31	6 24	11 17
15		2 45		7 26	12 18
16			6 33	8 27	13 19
17				9 29	14 20
18			7 36	9 30	15 21
19				10 31	16 22
20	1 65	4 51	8 38	11 32	17 23
21				12 33	18 24
22			10 40	12 34	19 26
23				13 35	20 27
24			11 43	14 36	21 28
25		6 56		15 38	22 29
26			12 45	16 39	22 30
27				16 40	23 31
28			14 47	17 41	24 32
29				18 42	25 32
30	4 74	8 61	15 49	19 43	26 34
31				20 44	27 35
32			17 51	21 45	28 36
33				21 46	29 37
34			18 53	22 47	30 38
35		11 66		23 48	31 39

TABLE IV -- (Continued)

PER CENT FAILURES IN SAMPLE	SIZE		OF		SAMPLE					
	10		20		50		100		1000	
36					20	55	24	49	32	40
37							25	50	33	41
38					21	57	26	51	34	42
39							27	52	35	43
40	8	81	15	70	23	59	28	53	36	44
41							29	54	37	45
42					24	61	29	55	38	46
43							30	56	39	47
44					26	63	31	57	40	48
45			18	74			32	58	41	49
46					28	65	33	59	42	50
47							34	60	43	51
48					29	67	35	61	44	52
49							36	62	45	53
50 *	13	87	22	78	31	69	37	63	46	54

* If percentage observed in sample exceeds 50 per cent, read 100 minus the percentage observed and subtract each confidence limit from 100.

TABLE V
UPPER OR MAXIMUM CONFIDENCE LIMIT (PER CENT) FOR
A SERIES OF TESTS WITHOUT A FAILURE

No. items tested without a failure	Confidence Levels for an interval type analysis				
	30 %	90%	95%	98 %	99 %
	Confidence Levels for a singular limit analysis				
	90%	95%	97 1/2%	99%	99 1/2 %
0					
1	90	95	98	99	100
2	68	78	84	90	93
3	54	63	71	79	83
4	44	53	60	68	73
5	37	45	52	60	65
6	32	39	46	54	59
7	28	35	41	48	53
8	25	31	37	44	48
9	23	28	34	40	44
10	21	26	31	37	41
11	19	24	29	34	38
12	17	22	27	32	36
13	16	21	25	30	33
14	15	19	23	28	31
15	14	18	22	27	30
16	13	17	21	25	28
17	13	16	19	24	27
18	12	15	19	23	26
19	11	15	18	22	24
20	11	14	17	21	23
25	8.8	11	14	17	19
30	7.4	9.5	12	14	16
35	6.4	8.2	10	12	14
40	5.6	7.2	8.8	11	12
45	5.0	6.4	7.9	9.8	11
50	4.5	5.8	7.1	8.8	10
75	3.0	3.9	4.8	6.0	6.8
100	2.3	3.0	3.6	4.5	5.2

If the last test, the tenth, for example, fails, Tables I, II, III, or IV must be used. It is not valid to ignore the last test and draw a conclusion from this table on the basis of nine tests without a failure. Furthermore, the experimenter must be prepared to use a single confidence limit type of analysis when he observes failures if he had planned to use this analysis in case he did not observe any failures.

For values not shown in the Table, the following formulae may be used:

For interval type analysis:

$$(1 - \text{conf. limit})^{\text{no. items tested}} = \frac{1 - \text{conf. level}}{2}$$

For single limit analysis:

$$(1 - \text{conf. limit})^{\text{no. items tested}} = 1 - \text{conf. level}$$

TABLE VI

NUMBER OF TESTS TO BE PERFORMED WITHOUT A FAILURE IN ORDER TO
DEMONSTRATE THE ACHIEVEMENT OF A SPECIFIED
MAXIMUM FAILURE RATE

Inasmuch as a specified maximum failure rate is, by the nature of its specification, a singular confidence limit, the confidence levels specified below refer to this type of analysis. The experimenter who proposes to use this table if he observes no failures, must also be prepared to express his results in the form of a maximum singular confidence limit if he does observe failures.

Specified Maximum Confidence Limit %	Confidence levels for a singular limit analysis				
	90 %	95 %	97 1/2 %	99 %	99 1/2 %
1	230	299	370	460	530
2	115	149	184	229	263
3	76	99	122	152	174
4	57	74	91	113	130
5	45	59	72	90	103
10	22	29	35	44	51
15	15	19	23	29	33
20	11	14	17	21	24
25	8	11	13	16	19

For example, an experimenter who wishes to be right on 95 per cent of his statements must perform 99 tests without a failure before he can say that the failure rate for an infinite number of the items is less than 3 per cent. (The formula is the same as for the single limit type analysis under Table V.) Obviously, if the true failure rate were 2 per cent, for example, he would most likely * get two failures in these 99 tests. Unfortunately, a sample of 99 with two failures could also have come from a population having a failure rate in excess of 3 per cent, the 95 per cent maximum confidence limit from Table II being 5 per cent for these results. The above table is therefore useful for an experimenter who believes that his failure rate is essentially zero but must demonstrate that it is less than a specified percentage. It is of no value to a man who believes his failure rate is just under the requirement. Even if he is correct in this belief, he should

* i. e. Two failures are more likely than any other number.

plan to test approximately 1000 samples before he can demonstrate this fact at 95 per cent confidence.

PART III

Purpose:

To define terms

a. In a series of perfectly controlled tests, the same answer should be obtained for each test. Inability or failure to duplicate the item under test and to reproduce the test conditions plus imperfections or lack of resolution in the measuring equipment make this impossible usually.

b. A good series of controlled tests is one in which the undesired variables are almost insignificant. Data might be as follows:

10.01
10.03
10.00
9.99

c. A poor series of controlled tests is one in which the undesired variables cause the data to be less consistent. Data might be as follows:

7.83
9.01
11.05
9.89

d. In order to study the effect of changing a characteristic of the item under test, the series of tests should be as good as possible, so that the change in performance, if this change exists, will not be hidden in a large spread of the data.

e. Assume data was obtained as follows:

9
12
10
9

The average or mean for an infinite number of these tests is desired but is not known. The average of the above series of four tests is 10 and is our best estimate of the average for an infinite number of tests. The more tests the better the estimate, generally. In other words, the sample average of 10 is the best estimate of the population average.

f. The goodness or badness of the series of tests may be measured in three ways:

(1) The range. This is the difference between the maximum and the minimum, or 3 in the above example. The range gets larger as the number of tests increases as the probability of the various undesired variables combining to form a very high or very low reading increases with the number of tests. The range, therefore is not the best measure of whether the experiment is good or bad, although it may be very useful, as noted in Par. n.

(2) The mean deviation or average deviation. The following table may be written:

Data	Best Estimate of Average	Deviation
9	10	-1
12		+2
10		0
9		-1

If the signs of the deviations are neglected and these deviations are averaged, the result is the mean deviation. This quantity appears attractive, but, to quote Villars, "... the distribution of this estimate is not amenable to simple mathematical manipulation ...", so it is not recommended.

(3) The standard deviation. The table is repeated

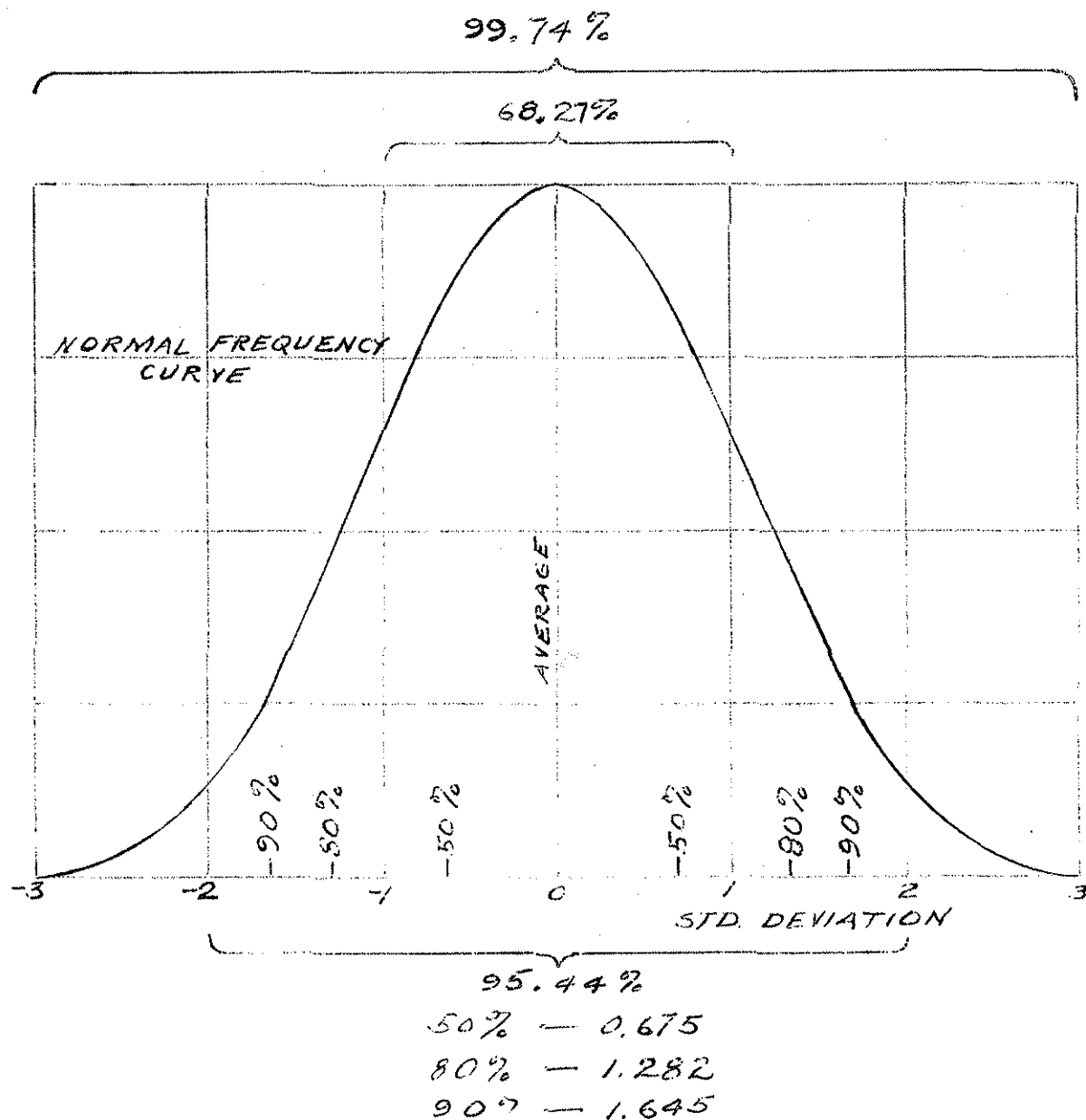
Data	Best Estimate of Average	Deviation	Deviation Squared
9	10	-1	1
12		+2	4
10		0	0
9		-1	1

The sum of the squares of deviation is 6. This is divided by 3, not 4, to give 2, the best estimate of the variance. The exact reason for this division by one less than the number of tests requires a study of statistics but it may be understood

intuitively by noting that the observed average value 10 which was used in the computation was only the best estimate of the true average and was too high or too low depending on how the data fell and this error in the observed average tended to make the deviations smaller than they should be. Hence, by dividing the sum of squares by one less than the number of tests, the best estimate of the variance is obtained. (Warning. Some authors of standard texts ignore this refinement, but for small numbers of tests it becomes significant.) The standard deviation is the square root of the variance, so the best estimate of the standard deviation for the above data is the square root of 2, or 1.4. This may be expressed as the relative standard deviation which is a percentage of the average or 1.4 divided by 10 equals 14 per cent. The standard deviation may also be called the standard error.

g. What is the utility of the best estimate of the standard deviation? Before investigating this question, let us consider for a moment the meaning of the standard deviation of an infinitely large number of items. The form of the distribution of the individual answers about the average value is not generally known, but the normal, or Gaussian distribution is generally preferred for the testing of commercially manufactured articles. A crude sketch of this distribution is shown on the next page. The percentages indicated thereon refer to the fraction of the total number of items tested which will fall between the indicated limits. For example, 68.27 per cent of all the items will have values between the average minus the standard deviation and the average plus the standard deviation. Similarly, 95.44 per cent of all the items will have values between the average minus twice the standard deviation and the average plus twice the standard deviation. It may be seen that 50 per cent of all the items will have values between the average minus about $2/3$ of a standard deviation and the average plus $2/3$ of a standard deviation. This 50 per cent point is referred to as the probable error and has been used considerably in the older literature.

h. A rough statement suitable for memorizing is that $2/3$ of all tests will have values lying within one standard deviation of the average, 95 per cent will have values lying within two standard deviations of the average, and practically all will lie within three standard deviations. The experimenter therefore looks at his best estimate of the standard deviation of 1.4 in the example posed in par. e. and his best estimate of the average of 10 and states that 95 per cent of a large number of tests will yield answers between $10 - 2.8$ and $10 + 2.8$, or between 7.2 and 12.8. Upon the basis of this statement, a large number of the items are made, but it is found that 95 per cent of these do not have values between 7.2 and 12.8. The experimenter



The percentages refer to the fraction of the total population which will have a value within the limits indicated for each percentage, i.e., 68.27 percent of the population has values between the average $\pm$ one std. deviation, or 80 percent of the population has values between the average $\pm$ (1.282)(std. deviation).

accepts this philosophically and retires to his laboratory licking his bruises and his performance rating. Next time he makes a similar statement a little hesitantly, but he makes it, and it also lands on him but hard. His faith in statistics has not been reinforced. What's wrong?

i. The difficulty probably lies not in the assumption of the normal distribution of the data; but in the fact that the best estimates of the average and the standard deviation will often be considerably in error when these best estimates are computed from a limited amount of data, as will become evident when confidence limits are computed for these quantities in Parts IV and VI of this memorandum.

j. Tables of tolerance factors for normal distributions have been computed by statisticians and may be found in reference h. of the Bibliography. The use of such tables is illustrated by reference to Table B and the example previously discussed in this Part. Table B is for 95 per cent confidence level. At this level of confidence, we may say that at least 95 per cent of a large number of tests will fall between our estimated average of ten plus or minus 6.37 times the estimated standard deviation of 1.4. This turns out to be from 1 to 19, which is not very definitive, to say the least, BUT IT IS THE BEST STATEMENT THAT CAN BE MADE AT 95 PER CENT CONFIDENCE!

k. Let us turn to another more reasonable example. Suppose that 20 flares have been tested, the average burning time for these items has been found to be 90.0 seconds, and the best estimate of the standard deviation has been found to be 4.7 seconds. What are the tolerance limits for 99 per cent of a large number of similar items? Referring to Table B, it may be said with 95 per cent confidence that at least 99 per cent of an infinite number of flares will have burning times between the average plus or minus (3.615) (4.7) seconds or between 73 and 107 seconds. At least 95 per cent of an infinite number of flares will have burning times between 90 plus or minus (2.752) (4.7) seconds or between 77 and 103 seconds, also at 95 per cent confidence. I like these tables because they enable the experimenter to make statements which are meaningful to the military and civilian authorities who have to make decisions relative to the probable utility of a weapon.

l. In many applications, the use of a singular tolerance limit is justified. In the preceding example, the experimenter may know that the maximum time that the flares might burn is of no interest but there is a keen interest in the minimum burning time. Again

referring to Table B he may say with 95 per cent confidence that at least 95 per cent of the flares will have a burning time of at least 79 seconds (90 minus 2.310×4.7).*

m. Even though the reader of a report is trained in statistics, which most readers are not, the specification of a standard deviation that is in reality only a best estimate of a standard deviation computed from a limited number of samples will almost invariably lead the reader to expect better results than he will realize because he visualizes the expected performance on the basis of accurate, not estimated, averages and standard deviations. The specification of tolerance limits avoids this difficulty, and expresses the results of the experiment in a form that is understandable to anyone who has mastered grade school arithmetic.

n. To go back to the range of the observed data, which was defined as the lowest observed value subtracted from the highest: As noted, the standard deviation is a better criterion and its estimate should be computed. If this is not practical in the field, a rough estimate of the standard deviation may be computed mentally after each test from the following table which is worth memorizing:

No. of tests in series	Divide range by this number to get rough estimate of standard deviation
4	2
6	2 1/2
10	3
15	3 1/2
30	4
100	5

o. In the original example, the range was 3 and there were 4 tests in the series. Dividing the 3 by 2 gives 1.5 as a rough estimate of the standard deviation, whereas the computation gave 1.4 as the best estimate. The above table also serves well in the evaluation of reported test results when the data is given in terms of the standard deviation. For example, if a report says there were 30 tests in the series and the average reading was 25.0 sec with a standard deviation of 4.0, the approximate range may be computed to be 16 so that it is most probable that the readings varied from 17 to 33. If the same data were given for a series

*This use of the table is not absolutely rigorous. It will be sufficiently accurate if the experimenter adds approximately 5 per cent to the value taken from the table. The computation would then be (90 minus $2.310 \times 1.05 \times 4.7$).

of 10 tests, the chances are that the range was 12 and the readings varied from 19 to 31. If the tests described in the report were to be duplicated, information of this type would be useful in setting the range of measuring equipment and/or the position of cameras.

p. There is a slick alternate method for computing the sum of squared deviations which gives the same answer but avoids the tedious computing of individual deviations. Add the individual readings in one column. Add the squares of the readings in a second column. Square the sum of the first column and divide by the number of tests. Subtract this number from the sum of the second column to get the sum of squared deviations. The example is worked below.

9	81	406
12	144	-400
10	100	<u>6</u>
9	81	
<u>40</u>	<u>406</u>	
x40		
4 <u>1600</u>		
400		

This method offers great advantages if a calculating machine is available as discussed in reference f. of the Bibliography.

NOTE: In using the tables on the next few pages, it is a conservative practice to add approximately 5 per cent to the values taken from the table when a single tolerance is being computed.

TABLE A -- TOLERANCE FACTORS FOR 90 PER CENT CONFIDENCE

These factors, expressed singularly or in $\pm$ form and multiplied by the best estimate of the standard deviation provide tolerances with respect to the observed average for the indicated percentage of an infinite population.

No. items Tested	Percentage of an Infinite Population for which				
	a Pair of Tolerances are Required For				
	75 %	90 %	95 %	99 %	99.9 %
	a Single Tolerance is Required For*				
	87.5 %	95 %	97.5 %	99.5 %	99.95 %
2	11.407	15.978	18.800	24.167	30.227
3	4.132	5.847	6.919	8.974	11.309
4	2.932	4.166	4.943	6.440	8.149
5	2.454	3.494	4.152	5.423	6.879
6	2.196	3.131	3.723	4.870	6.188
7	2.034	2.902	3.452	4.521	5.750
8	1.921	2.743	3.264	4.278	5.446
9	1.839	2.626	3.125	4.098	5.220
10	1.775	2.535	3.018	3.959	5.046
11	1.724	2.463	2.933	3.849	4.906
12	1.683	2.404	2.863	3.758	4.792
13	1.648	2.355	2.805	3.682	4.697
14	1.619	2.314	2.756	3.618	4.615
15	1.594	2.278	2.713	3.562	4.545
16	1.572	2.246	2.676	3.514	4.484
17	1.552	2.219	2.643	3.471	4.430
18	1.535	2.194	2.614	3.433	4.382
19	1.520	2.172	2.588	3.399	4.339
20	1.506	2.152	2.564	3.368	4.300
21	1.493	2.135	2.543	3.340	4.264
22	1.482	2.118	2.524	3.315	4.232
23	1.471	2.103	2.506	3.292	4.203
24	1.462	2.089	2.489	3.270	4.176
25	1.453	2.077	2.474	3.251	4.151
26	1.444	2.065	2.460	3.232	4.127
27	1.437	2.054	2.447	3.215	4.106
28	1.430	2.044	2.435	3.199	4.085
29	1.423	2.034	2.424	3.184	4.066
30	1.417	2.025	2.413	3.170	4.049
40	1.370	1.959	2.334	3.066	3.917
50	1.340	1.916	2.284	3.001	3.833
100	1.275	1.822	2.172	2.854	3.646
200	1.234	1.764	2.102	2.762	3.529
300	1.217	1.740	2.073	2.725	3.481
400	1.207	1.726	2.057	2.703	3.453
500	1.201	1.717	2.046	2.689	3.434
1000	1.185	1.695	2.019	2.654	3.390
∞	1.150	1.645	1.960	2.576	3.291

* See Note, page 29.

TABLE B -- TOLERANCE FACTORS FOR 95 PER CENT CONFIDENCE

These factors, expressed singularly or in $\pm$ form and multiplied by the best estimate of the standard deviation provide tolerances with respect to the observed average for the indicated percentage of an infinite population.

No. items Tested	Percentage of an Infinite Population for which a pair of Tolerances are Required for				
	75%	90%	95%	99%	99.9%
	a Single Tolerance is Required for*				
	87.5%	95%	97.5%	99.5%	99.95%
2	22.858	32.019	37.674	48.430	60.573
3	5.922	8.380	9.916	12.861	16.208
4	3.779	5.369	6.370	8.299	10.502
5	3.002	4.275	5.079	6.634	8.415
6	2.604	3.712	4.414	5.775	7.337
7	2.361	3.369	4.007	5.248	6.676
8	2.197	3.136	3.732	4.891	6.226
9	2.078	2.967	3.532	4.631	5.899
10	1.987	2.839	3.379	4.433	5.649
11	1.916	2.737	3.259	4.277	5.452
12	1.858	2.655	3.162	4.150	5.291
13	1.810	2.587	3.081	4.044	5.158
14	1.770	2.529	3.012	3.955	5.045
15	1.735	2.480	2.954	3.878	4.949
16	1.705	2.437	2.903	3.812	4.865
17	1.679	2.400	2.858	3.754	4.791
18	1.655	2.366	2.819	3.702	4.725
19	1.635	2.337	2.784	3.656	4.667
20	1.616	2.310	2.752	3.615	4.614
21	1.599	2.286	2.723	3.577	4.567
22	1.584	2.264	2.697	3.543	4.523
23	1.570	2.244	2.673	3.512	4.484
24	1.557	2.225	2.651	3.483	4.447
25	1.545	2.208	2.631	3.457	4.413
26	1.534	2.193	2.612	3.432	4.382
27	1.523	2.178	2.595	3.409	4.353
28	1.514	2.164	2.579	3.388	4.326
29	1.505	2.152	2.564	3.368	4.301
30	1.497	2.140	2.549	3.350	4.278
40	1.435	2.052	2.445	3.213	4.104
50	1.396	1.996	2.379	3.126	3.993
100	1.311	1.874	2.233	2.934	3.748
200	1.258	1.798	2.143	2.816	3.597
300	1.236	1.767	2.106	2.767	3.535
400	1.223	1.749	2.084	2.739	3.499
500	1.215	1.737	2.070	2.721	3.475
1000	1.195	1.709	2.036	2.676	3.418
∞	1.150	1.645	1.960	2.576	3.291

*See Note, page 29.

TABLE C--TOLERANCE FACTORS FOR 99 PER CENT CONFIDENCE

These factors, expressed singularly or in $\pm$ form and multiplied by the best estimate of the standard deviation provide tolerances with respect to the observed average for the indicated percentage of an infinite population.

No. items Tested	Percentage of an Infinite Population for which a pair of tolerances are required for				
	75 %	90 %	95 %	99 %	99.9 %
	a single tolerance is required for*				
	87.5 %	95 %	97.5 %	99.5 %	99.95 %
2	114.363	160.193	188.491	242.300	303.054
3	13.378	18.930	22.401	29.055	36.616
4	6.614	9.398	11.150	14.527	18.383
5	4.643	6.612	7.855	10.260	13.015
6	3.743	5.337	6.345	8.301	10.548
7	3.233	4.613	5.488	7.187	9.142
8	2.905	4.147	4.936	6.468	8.234
9	2.677	3.822	4.550	5.966	7.600
10	2.508	3.582	4.265	5.594	7.129
11	2.378	3.397	4.045	5.308	6.766
12	2.274	3.250	3.870	5.079	6.477
13	2.190	3.130	3.727	4.893	6.240
14	2.120	3.029	3.608	4.737	6.043
15	2.060	2.945	3.507	4.605	5.876
16	2.009	2.872	3.421	4.492	5.732
17	1.965	2.808	3.345	4.393	5.607
18	1.926	2.753	3.279	4.307	5.497
19	1.891	2.703	3.221	4.230	5.399
20	1.860	2.659	3.168	4.161	5.312
21	1.833	2.620	3.121	4.100	5.234
22	1.808	2.584	3.078	4.044	5.163
23	1.785	2.551	3.040	3.993	5.098
24	1.764	2.522	3.004	3.947	5.039
25	1.745	2.494	2.972	3.904	4.985
26	1.727	2.469	2.941	3.865	4.935
27	1.711	2.446	2.914	3.828	4.888
28	1.695	2.424	2.888	3.794	4.845
29	1.681	2.404	2.864	3.763	4.805
30	1.668	2.385	2.841	3.733	4.768
40	1.571	2.247	2.677	3.518	4.493
50	1.512	2.162	2.576	3.385	4.323
100	1.383	1.977	2.355	3.096	3.954
200	1.304	1.865	2.222	2.921	3.731
300	1.273	1.820	2.169	2.850	3.641
400	1.255	1.794	2.138	2.809	3.589
500	1.243	1.777	2.117	2.783	3.555
1000	1.214	1.736	2.086	2.718	3.472
∞	1.150	1.645	1.960	2.576	3.291

*See Note, page 29.

PART IV

Purpose:

To determine the confidence limits for an average of a series of tests.*

a. If a series of tests is performed, the average of the answers is the best estimate of the average for an infinite number of tests. The estimate is better if there are a large number of tests in the series, and the estimate is better if the variance and standard deviation are low. The experimenter knows that the average for the series is the best possible estimate of the true average, but it is necessary to assign confidence limits to the average before any statement of how good the estimate is may be made.

b. To review:

- (1) The deviation of a particular test in a series is the difference between the answer for that test and the average for the series.
- (2) The sum of squared deviations is self-explanatory.
- (3) The variance is the sum of squared deviations divided by one less than the number of tests in the series.
- (4) The standard deviation is the square root of the variance.

c. A problem will be proposed and solved. Assume that the penetration of a shaped charge is measured in inches as follows:

Data	Avg.	Dev.	Dev. sq.	Variance	Std. Dev.
55	50	+5	25	19	4.36
47		-3	9		
48		-2	4		
			38		

* It is assumed that the population sampled has a normal, or Gaussian distribution in this and all succeeding parts.

The best estimate of the true average is 50. If we had selected three different samples, a different average for the second series would be obtained. If this sampling and testing in series of three were continued, a whole set of averages would be obtained. Since these could be tabulated, we could average them, compute deviations, square them and find a variance for the average. This large amount of testing is not required, however, as it may be shown that the variance for the average is merely the variance for an individual test divided by the number of tests in a series. For our example, 19 divided by 3 gives 6.33 as the variance for averages, each of which is computed from a series of three tests. The standard deviation for this type of average is $\sqrt{6.33}$ or 2.52.

d. We may say that the true average lies between

$$50 \pm (t) (\text{std. dev. of average}) \text{ or } 50 \pm t \sqrt{\frac{\text{var. for tests}}{\text{no. of tests}}}$$

t is taken from the Table of Students t Distribution. The degrees of freedom is the number we divided the sum of squared deviations by, or 2 in this example. (One less than the number of tests in the series.) If we require 95 per cent confidence, $t = 4.30$. The confidence interval limits for the average are therefore

$$50 \pm (4.30) (2.52) = 50 \pm 10.9$$

or 39.1 to 60.9 is the confidence interval, for the average.

If we can afford to be right only 80 per cent of the time, $t = 1.886$ and the limits for the average are therefore

$$50 \pm (1.886) (2.52) = 50 \pm 4.8$$

or 45.2 to 54.8 is the confidence interval, for the average.

e. It is interesting to note that the confidence interval for 95 per cent confidence covers a zone that is somewhat greater than the total spread of the three observations. A study of the t table shows that t decreases very rapidly for the first ten tests in a series but the next ten are of much less value in this respect.

f. If several series of tests have been performed, each with some characteristic of the item under test slightly different, there is a means whereby the number of degrees of freedom may be increased

and a smaller confidence interval obtained for the average for each series. This technique is described in Part V, paragraph e, and should be utilized whenever applicable.

g. If the experimenter is only interested in how high or how low the true average may be, he can make a more precise statement employing only a single confidence limit and still retain his confidence level. In the example above, the maximum confidence limit for the true value of the average is

$$50 + (2.92)(2.52) = 57.4$$

The formal statement is therefore: "The average value for three tests was 50 inches and the writer is confident (95 per cent) that the true average for an infinite number of items made to identical specifications will be less than 57.4 inches."

h. A similar but separate statement could have been made to the effect that the true value of the average would be greater than 42.6 inches, but the statements cannot be combined into a single statement without lowering the confidence to a 90 per cent level.

i. The t table has been appropriately labeled for the determination of both a confidence interval and a singular confidence limit. The experimenter may use either an interval statement or a singular confidence limit statement, as he sees fit, for each series of tests.

j. Data of this type will not fit a normal distribution exactly as the normal distribution extends from $-\infty$ to $+\infty$ and our data cannot be less than zero. If the average is of the same order of magnitude as the standard deviation of the average, the confidence interval may extend into the negative region, an obvious impossibility. A statistician should be consulted as there are many other types of distributions, and he can probably fit the data to one of these.

k. Although the stating of an average for a series of tests and the indication of a confidence interval for that average is standard practice, I believe that the tolerance limits discussed in par. j of Part III are easier to compute and provide the reader with a more usable type of information.

STUDENT'S t DISTRIBUTION

Degrees of Freedom	Confidence level for an interval type of analysis and for the Student's t Test						
	99.9%	99%	98%	95%	90%	80%	70%
Confidence level for a singular confidence limit (Do not use for Student's t Test except as described in Part V, par. z.)							
	99.95%	99.5%	99%	97.5%	95%	90%	85%
1	636.62	63.657	31.821	12.706	6.314	3.078	1.963
2	31.60	9.925	6.965	4.303	2.920	1.886	1.386
3	12.94	5.841	4.541	3.182	2.353	1.638	1.250
4	8.61	4.604	3.747	2.776	2.132	1.533	1.190
5	6.86	4.032	3.365	2.571	2.015	1.476	1.156
6	5.96	3.707	3.143	2.447	1.943	1.440	1.134
7	5.41	3.499	2.998	2.365	1.895	1.415	1.119
8	5.04	3.355	2.896	2.306	1.860	1.397	1.108
9	4.78	3.250	2.821	2.262	1.833	1.383	1.100
10	4.59	3.169	2.764	2.228	1.812	1.372	1.093
11	4.44	3.106	2.718	2.201	1.796	1.363	1.088
12	4.32	3.055	2.681	2.179	1.782	1.356	1.083
13	4.22	3.012	2.650	2.160	1.771	1.350	1.079
14	4.14	2.977	2.624	2.145	1.761	1.345	1.076
15	4.07	2.947	2.602	2.131	1.753	1.341	1.074
16	4.02	2.921	2.583	2.120	1.746	1.337	1.071
17	3.97	2.898	2.567	2.110	1.740	1.333	1.069
18	3.92	2.878	2.552	2.101	1.734	1.330	1.067
19	3.88	2.861	2.539	2.093	1.729	1.328	1.066
20	3.85	2.845	2.528	2.086	1.725	1.325	1.064
21	3.82	2.831	2.518	2.080	1.721	1.323	1.064
22	3.79	2.819	2.508	2.074	1.717	1.321	1.063
23	3.77	2.807	2.500	2.069	1.714	1.319	1.060
24	3.75	2.797	2.492	2.064	1.711	1.318	1.059
25	3.73	2.787	2.485	2.060	1.708	1.316	1.058
26	3.71	2.779	2.479	2.056	1.706	1.315	1.058
27	3.69	2.771	2.473	2.052	1.703	1.314	1.057
28	3.67	2.763	2.467	2.048	1.701	1.313	1.056
29	3.66	2.756	2.462	2.045	1.699	1.311	1.055
30	3.65	2.750	2.457	2.042	1.697	1.310	1.055
∞	3.29	2.576	2.326	1.960	1.645	1.282	1.036

PART V

Purpose:

To determine whether a deliberate change in the item under test changes its performance.

a. If a series of tests were perfect, and the answers identical, even the slightest change in the item would be immediately apparent. Tests are not perfect and it is therefore more difficult to tell if an apparent change in performance is caused by the deliberate change or the inherent perversity of the experiment. The "Student's t Test" is often of assistance and its use will be explored.

b. To review:

- (1) The deviation of a particular test in a series is the difference between the answer for that test and the average for the series.
- (2) The sum of squared deviations is self-explanatory.
- (3) The variance is the sum of squared deviations divided by one less than the number of tests in the series.
- (4) The standard deviation is the square root of the variance.

c. Assume the penetration of a shaped charge is measured in inches as follows for a series of three tests with a certain liner. The composition of the liner is changed from Alloy A to Alloy B and the series is repeated.

A		B	
	55		55
	47		64
	48		64
Total	150	Total	183
Average	50	Average	61

It appears that Alloy B is better, but maybe the data would have spread out that much if the first series of tests had consisted

of six tests and the liner wasn't changed. The liners, being different, are referred to as treatments.

d. The problem will be worked and explained simultaneously. (The problem and the solution were taken partially from Dr. Villar's notes for a forthcoming book on small sample statistics which has since been published and is listed in the Bibliography.) We know that the best estimate of the average for Treatment A is 50 and for Treatment B is 61. The difference, which shall be designated x , is 11. If we were to run a number of identical series of tests, each series to consist of three tests, we know that the individual averages (each a best estimate for its series) would not be identical, although these averages would group closer than individual answers. Our question resolves itself into an inquiry as to whether a difference in averages of 11 is worth getting excited over.

e. First, the perversity of the test must be considered. This is best measured by computing best estimates for the variance and standard deviation. This poses a problem. The results cannot be combined about a gross average of 55 1/2 as we suspect the two series are not identical. Neither can a very accurate estimate be made for a series of only three tests. However, our change in liner should not have affected the degree of perversity of the test so the two series can be combined to advantage as follows:

Data	Average	Deviation	Deviation Squared
55 } 47 } 48 }	50	{ +5 -3 -2	25 9 4
55 } 64 } 64 }	61	{ -6 +3 +3	36 9 9
<u>Sum of Squared Deviations</u>			<u>92</u>

NOTE: It is not essential that both series consist of an equal number of tests.

f. For one series of tests, the sum of squared deviations was divided by one less than the number of tests in the series to get the best estimate of the variance. With two series, we divide by a number which is the sum of two numbers. Each number

is the number that would have been used for its particular series alone. In this case, two would have been used for each series so two plus two equals four and the sum of squared deviations is divided by four. If the second series had had four tests, the number for this series would have been three, and two plus three equals five. (If the data had been taken from three series of three tests each with three different liners, the number would have been two plus two plus two equals six, but that may be skipped for the moment.)

g. Dividing the sum of squared deviations of 92 by 4 gives 23 which is the best estimate of the variance for the type of experiment we are performing. The best estimate of the standard deviation is the square root of 23, or 4.8.

h. To return to the variance of 23. If the variance for each individual test is known, the variance for the difference between averages of two different series of tests may be estimated.

$$\begin{aligned} \text{Variance for the} & & \text{Variance for} & \left(\frac{1}{m} + \frac{1}{n} \right) \\ \text{difference in averages} & = & \text{each test} & \\ & & & \\ & = & 23 & \left(\frac{1}{3} + \frac{1}{3} \right) \\ & & & \\ & = & 15.33 & \end{aligned}$$

where m and n are the respective numbers of tests in each series.

i. The square root of this variance of 15.33 is 3.92, which is the best estimate of the standard deviation of the difference in averages of different identical series of tests with three tests per series of the type under discussion.

j. If we define

$$\text{computed } t = \frac{\bar{x}}{\text{Std. dev. for diff. in avgs.}} = \frac{11}{3.92} = 2.81$$

we may refer to the t graph at the end of this part. This graph is a plot of the t table at the end of Part IV. Either may be used; the graph obviates the need for interpolation, but the table is more precise. The degrees of freedom is a term that goes with our estimate of the variance for each test and is the number we divided the sum of squared deviations by to get this variance. For our example, we have four degrees of freedom and we use the curve corresponding to four. We find the computed value of t, namely 2.81, on the vertical axis, move across to the curve,

and read the abscissa of this point on the horizontal scale at the bottom. For our example, the value is slightly less than 0.05. This is the probability of our getting a difference of 11 or more in two averages of three tests each if the alloy of the liners had not been changed. Stated another way, in only five times in a hundred, on the average, would we observe such a great difference in averages if there were no difference between Alloy A and Alloy B.

k. Therefore it seems reasonable to state that there is a difference between Alloy A and Alloy B at 95 per cent confidence. We can expect to be wrong in only 5 per cent of such statements. To simplify the use of the t graph, the appropriate confidence level values have been filled in along the top of the graph. The experimenter computes his value of t, enters the graph using the appropriate degrees of freedom curve, and determines the confidence level at which he can proclaim that one treatment surpasses the other.

l. A rose is a rose, etc., according to Stein, and in statistics significance has significance. To the uninitiated experimenter a significant difference in two treatments, such as the liners in the shaped charges, is a worthwhile difference in terms of performance or economic saving or some other criterion. Not so in the phraseology of some statisticians. A significant change is any difference which can be proclaimed at 95 per cent confidence. For our example, our test demonstrates that there is a significant difference in the two alloys. If our experiment were less perverse, or if we could have afforded more items in each series, our experiment would have been more sensitive and might have been able to claim a significant difference even though the observed difference in the performance of the two alloys was actually less, in terms of inches of penetration.

m. If the computed t were greater than 4.60, the superiority of one liner with respect to the other would have been proclaimed at 99 per cent confidence and this would have been defined statistically as a highly significant difference. Likewise, if the computed t were greater than 8.61, the difference would be proclaimed at 99.9 per cent confidence, and the difference would be considered as very highly significant! This rather arbitrary phraseology is not objectionable if the reader of a technical report understands what these terms mean. For example, if the computed t had equaled 2.20, some statisticians might say that "no significant difference was demonstrated." I might say that, "Alloy B is better than Alloy A, at better than 90 per cent confidence." The statements appear contradictory, but they are not because the statistician prefers

not to commit himself unless he can be correct in at least 95 per cent of his statements. His statement that no significant difference was demonstrated is his way of declining to express an opinion, but it may be misleading to a reader who is not versed in the phraseology. For this reason, I do not recommend using the terms significant, highly significant and very highly significant as synonyms for certain degrees of confidence. State the apparent direction of the superiority, and indicate the confidence level.

n. For a little practice in applying the Student t test, another example is proposed as follows:

Series A

71
67
33
79
42
Average 58.4

Series B

67
86
Average 76.5

Was the difference in averages of 18.1 a reliable indication of the superiority of Treatment B over Treatment A?

Data	Average	Deviation	Deviation Squared
$\left. \begin{array}{l} 71 \\ 67 \\ 33 \\ 79 \\ 42 \end{array} \right\}$	58.4	$\left\{ \begin{array}{l} +12.6 \\ +8.6 \\ -25.4 \\ +20.6 \\ -16.4 \end{array} \right.$	$\begin{array}{r} 158.76 \\ 73.96 \\ 645.16 \\ 424.36 \\ 268.96 \end{array}$
$\left. \begin{array}{l} 67 \\ 86 \end{array} \right\}$	76.5	$\left\{ \begin{array}{l} -9.5 \\ +9.5 \end{array} \right.$	$\begin{array}{r} 90.25 \\ 90.25 \end{array}$
			1,751.70

Divide by five (four plus one) to get a variance of 350.34 for each test. The variance for the difference in the averages equals $350.34(1/5 + 1/2)$ which equals 245.238. The standard deviation for the differences is the square root of this or 15.7. The difference in averages is 18.1. Computed $t = \frac{18.1}{15.7} = 1.15$ for 5 degrees

of freedom. This corresponds to a significance level of 0.3, that is, there is a probability of 0.3 of getting this great a difference in the two averages if there were no difference in the two treatments. We can conclude that Treatment B is better but we can expect to be right in only 70 per cent of such statements. More tests are needed when the data scatters this badly.

o. To return to the example in par. c of this part, suppose the experimenter works on a 99 per cent confidence level. What difference in the two averages must exist for him to make a positive statement?

$$\text{Tabular } t = \frac{\text{Min } x}{\text{std. dev. for diff. in avg.}}$$

$$4.604 = \frac{\text{Min } x}{3.92}$$

$$\text{Min } x = 18.05$$

If the difference in averages is less than 18.05, a 99 per cent-confidence man will not state definitely that Alloy B is better than Alloy A. He may perform more tests per series. This will give him a better estimate of the variance per test and more degrees of freedom so a smaller difference may be "highly significant." To illustrate this, the experimenter wants to know how many tests per series he must perform before the present difference in average of 11 (assuming this difference does not change appreciably) may be considered "highly significant."

$$\text{Tabular } t = 4.604 = \frac{11}{\text{std. dev. for diff. in avg. (Max.)}}$$

Maximum standard deviation for difference in average = 2.39, variance for this standard deviation = 5.72.

$$5.72 = 23 \left(\frac{1}{n} + \frac{1}{n} \right)$$

$$n = 8+, \text{ or } 9 \text{ tests per series.}$$

This would appear to be conservative because he will then have 16 degrees of freedom and the tabular t will be lower, actually 2.921. A person could assume that the variance estimate of 23 and the average difference of 11 will not change and solve by trial and error

for the minimum number of tests per series which will indicate at the 99 per cent confidence level that there is a difference in the two treatments. Unfortunately, this technique will reduce to 50 per cent the probability that a real difference of 11 inches in the two treatments will be detected at the 99 per cent confidence level.

Actually, the experimenter may have to perform more, not less than 9 tests per series to provide a high assurance that he will detect the required difference if it does exist. The solution of this problem of determining the number of tests per series is not elementary, and I would recommend that the reader consult a statistician, or if this is impracticable, an intensive study of pages 17 to 26 of reference k of the Bibliography will be of value. It is also suggested that the experimenter should determine from his knowledge of the application of the item just what improvement he considers worthwhile detecting. In our example, it may be that an improvement of 5 inches in the average penetration would be worthwhile; this will be difficult to establish: if an improvement must consist of 20 inches before it is of interest, this can be easily detected. The choice of 11 inches merely because it was observed in the initial tests is not a logical method of choosing the difference in performance that is of interest.

p. If the averages do not differ by a "significant" amount, the two series of tests are often lumped together by statisticians if this is desirable.

q. To return to the equation for the variance of the difference in averages:

$$\text{Variance for diff. in avg.} = \text{var. per test} \left(\frac{1}{m} + \frac{1}{n} \right)$$

It is desirable to have $m = n$ for the most efficient comparison of two series. The expression $\left(\frac{1}{m} + \frac{1}{n} \right)$ should be small so that the computed t will be large. For a given total number of tests $\left(\frac{1}{m} + \frac{1}{n} \right)$ will be least when $m = n$.

r. In some cases $m = n$ is not practical but the formulae are still applicable. The degree of freedom is equal to $m - 1$ plus $n - 1$.

s. If one series of tests consists of one test only, the $\left(\frac{1}{m} + \frac{1}{n}\right)$ term reduces to $\left(1 + \frac{1}{n}\right)$ and the variance per test must be taken from the series of n tests only and the degrees of freedom is therefore n-1.

t. If one series of tests consists of n tests and the other value is assumed to be precise and representative of a very large number, or is a requirement, the $\left(\frac{1}{m} + \frac{1}{n}\right)$ term reduces to $\frac{1}{n}$ and the variance per test must be taken from the series of n tests only. Therefore the degrees of freedom is n-1.

u. There is another approach that may be utilized to indicate the difference between two treatments, such as the use of Alloy A and Alloy B in the shaped charge liners, in the example in par. c. of this part. The best estimate of the difference in the performance of heads with the two types of liners is 11 inches. What are the confidence limits on this difference in performance of 11 inches? As noted in par. i. of this part, the best estimate of the standard deviation of the difference in these averages is 3.92. For four degrees of freedom and 80 per cent confidence, the tabular t = 1.533. The confidence limits for the difference are therefore

$$11 \pm (3.92) (1.533) = 11 \pm 6.01$$

The lower and upper 80 per cent confidence limits are therefore 5.0 and 17.0 for the difference in performance of the two liners.

If 95 per cent confidence is required, the tabular t = 2.776 and the limits are

$$11 \pm (3.92) (2.776) = 11 \pm 10.9$$

The lower and upper 95 per cent confidence limits are therefore 0.1 and 21.9. In some respects, this method of comparing two treatments may be more informative than merely saying that there is a difference at a specified confidence level.

v. It was noted in par. m of this part that a 99 per cent-confidence man would refrain from making any statement about the superiority of Alloy B over Alloy A on the grounds that the data was not sufficient to convince him. What will happen if we use the technique of par. u at 99 per cent confidence level? The tabular t

equals 4.604 so the limits are

$$11 \pm (3.92) (4.604) = 11 \pm 18.1$$

The upper confidence limit is now 29.1; the lower confidence limit is -7.1! Small wonder that a 99 per cent-confidence man refuses to commit himself. He could say, "Yes, I will state that Alloy B is superior to Alloy A by a margin of from minus 7.1 inches to plus 29.1 inches." Such an oblique statement is hardly suitable for a formal report, however, so it is preferable to make a more definitive statement at a lower level of confidence. The experimenter should always indicate his confidence level with his statements.

w. Returning to Part IV in which the technique of determining confidence limits for the average of a series of tests was discussed, it may be shown that a better estimate of the variance per test can be obtained if several series of tests have been performed, even though a different treatment was employed for each series. The perversity of the experiment should remain the same for each series even though the liners are made of different alloys, and it is the perversity of the experiment, as expressed by the variance, that leads to differences between sample averages and the true average for a large number of tests. Therefore, if the series of tests proposed in Part IV were the first series of tests of the two proposed in par. c. of this part, the 95 per cent confidence limits for the average of the first series may be recomputed to be

$$50 \pm 2.776 \sqrt{\frac{23}{3}} = 50 \pm 7.7$$

or 42.3 to 57.7

It is of interest to note that although the best estimate of the variance is a larger number (23 compared to 19), this is more than counterbalanced by the additional certainty obtained by more tests and hence more degrees of freedom.

x. "Why didn't he mention the use of a singular confidence limit in par. u? I'm only interested in the minimum (or maximum) amount by which Alloy B may be superior to Alloy A, so why should I bother with an interval technique?" Okay, you can use a singular confidence limit technique, IF and only IF you are really interested in the superiority of Alloy B over Alloy A. BUT you must know that you are interested in this BEFORE you analyze the data. You cannot look at the data to see which Alloy came out ahead and then decide

which one you are interested in. If you wish to utilize the results of the test to decide which treatment is superior, then you must use an interval technique as described in par. u.

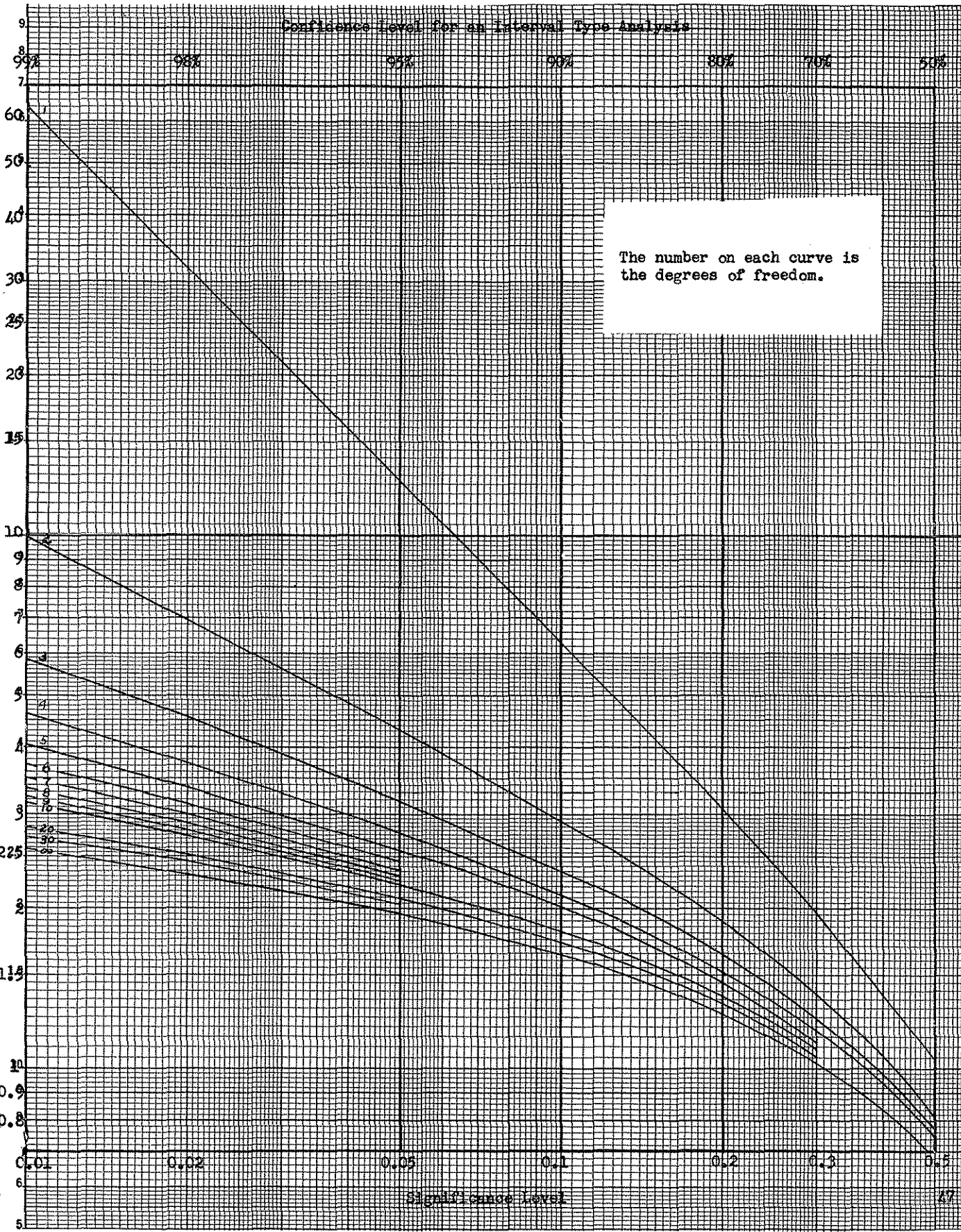
y. Look at it this way. Suppose the experimenter decides in advance that he will subtract the second average from the first average, regardless of how the results come out. In our example, the difference would then be -11 inches. The lower 95 per cent confidence limit would be -21.9 inches, and the upper 95 per cent confidence limit be -0.1 inches; the 95 per cent interval would extend from -21.9 inches to -0.1 inches. (I suppose the report would say that Alloy A was superior to Alloy B by minus 11 inches!) But suppose that Alloy A had an average of 11 inches greater than that for Alloy B? The lower limit is now +0.1 inches and upper limit is +21.9 inches. Note how the absolute critical values of the upper and lower limits have switched on us, depending upon which Alloy came out ahead. Which limit is the upper and which is the lower depends upon which treatment comes out ahead, and this dependency makes it impossible to use the singular confidence limit type of analysis.

z. Let us assume that the experimenter knew in advance that he was only interested in the possible superiority of Alloy B over Alloy A, since Alloy B is much more expensive and more difficult to fabricate. He could then decide in advance to subtract the average for Alloy A from Alloy B, regardless of how the results came out. In our example, his difference would be a plus 11 inches. After he has computed a value of t in the usual manner, he can compare it with the tabular t values taken from the singular limit confidence levels shown in the table. He can then state the level at which he is confident that Alloy B is superior to Alloy A. (If the difference came out negative, or the confidence level is less than 50 per cent, he will state that the superiority of Alloy B was not demonstrated, but he cannot make any statement relative to the possible superiority of Alloy A.) He can compute a minimum or maximum limit (but not both) relative to the superiority of Alloy B over Alloy A. Typical statements for our example might be: "The writer is 95 per cent confident that Alloy B is superior to Alloy A by at least 2.6 inches of penetration,"¹ or "The writer is 95 per cent confident that the superiority of Alloy B over Alloy A will not exceed 19.4 inches."²

¹ $11 - 3.92 \times 2.132$

² $11 + 3.92 \times 2.132$

Confidence Level for an Interval Type Analysis



Significance Level

aa. If a manufacturing facility proposes the relaxation of a fabricating tolerance in order to reduce the percentage of rejects, or the substitution of propellant material of inferior physical characteristics, the experimenter is not interested in whether the proposed change will yield better results; in this example he knows that it won't. He is only interested if the performance will be impaired. He agrees to subtract the average performance for several items utilizing the proposed change from the average performance for the items utilizing the original design and then performs a singular limit type analysis. He would probably solve for a maximum confidence limit and say, "The writer is 95 per cent confident that the decrease in the mean effective pressure caused by the utilization of the inferior propellant as recommended by the loading plant will not exceed _____ pounds per square inch." It is interesting to note that even if the difference in averages turns out to be zero or slightly negative, the maximum confidence limit will still exist and may still be positive, since the maximum confidence limit is the observed difference plus the tabular t times the standard deviation for the difference in the averages.

bb. To summarize, if the determination of the superiority or inferiority (but not both) of one specified treatment with respect to another is the sole problem, a singular limit type analysis is more **informative** and should be employed. If the determination of which treatment is superior is the problem, the interval type analysis discussed in paragraphs a through w should be employed.

PART VI

Purpose:

To determine the confidence limits for the standard deviation of a series of tests.

a. After the experimenter computes a best estimate of the standard deviation of a series of tests, he may ask himself how good this estimate is. In order to answer this question, he may wish to compute the confidence interval or a singular confidence limit for the standard deviation. Generally speaking he will be interested in only the maximum confidence limit, although in the case of a weapon which depends upon dispersion for its effectiveness, such as a shotgun or a barrage technique, he may wish to compute a confidence interval or even a minimum confidence limit.

b. At the end of the Part is a table of B factors¹ which are to be multiplied by the standard deviation. If the experimenter wishes to know the maximum confidence limit for the standard deviation, he selects his confidence level and enters the table on the line opposite the appropriate degree of freedom and selects the B2 factor, which he then multiplies by the standard deviation to obtain the maximum confidence limit for his standard deviation. He can then state that the true standard deviation for an infinite number of items will be less than this value. If he were interested in a minimum confidence limit he would have selected the B1 factor.

c. If the experimenter were interested in a confidence interval he would have utilized both the factors to obtain the upper and lower limits for his interval but he would have entered the table in a different column for the same confidence level.

d. The number of degrees of freedom is one less than the number of items tested in a single series of tests. Using the example from Part IV, the best estimate of the standard deviation was found to be the square root of 19, or 4.36. Three items were tested so there are two degrees of freedom. At 95 per cent confidence, the experimenter can say that the standard deviation is less than 19.2 (4.406×4.36). If he were interested in a confidence interval

¹This table was prepared by Eleanor G. Crow, who has simplified the classical technique by performing the square root of the quotient of the degrees of freedom and the value of x^2 .

he could say at 95 per cent confidence that the standard deviation is between 2.27 and 27.4 (0.521×4.36 and 6.285×4.36).

e. The interval computed above is quite broad and even the maximum confidence limit is unencouragingly high. This serves to emphasize the fact that a good estimate of the standard deviation can not be obtained with only three tests in a series. The statements become much more precise as the number of tests in the series is increased. If the experimenter had performed the two series of tests as described in Part V, he has an estimate of the standard deviation of 4.80 and four degrees of freedom to work with. At 95 per cent confidence, he obtains a maximum confidence limit of 11.4 or a confidence interval of from 2.87 to 13.8 for the standard deviation. A general rule for the computation of the number of degrees of freedom is to subtract the number of different series of tests (in this case two series) from the grand total of the number of items tested (in this case six items). The justification for the apparent shenanigan of combining two series of tests which are considered non-identical lies in the assumption that the perversity, or tendency of the data to spread, is the same for each series of tests, since only the alloy of the liner was changed, and this should not affect the degree of nonreproducibility of a series of tests. The standard deviation is a measure of the perversity of an experiment so it and its confidence limits can better be computed from both series of tests.

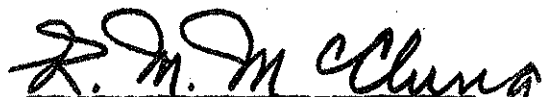

 R. M. McCLUNG, Head
 Development Division

TABLE OF B FACTORS

Factors for Determining Confidence Limits or
Intervals for a Standard Deviation

Degrees of Freedom	Confidence levels for an interval type analysis					
	90 %		95 %		98 %	
	Confidence levels for a singular limit analysis					
	95 %		97.5 %		99 %	
	B1	B2	B1	B2	B1	B2
1	0.510	15.952	0.446	32.226	0.388	252.377
2	0.578	4.406	0.521	6.285	0.466	9.975
3	0.620	2.919	0.566	3.723	0.514	5.108
4	0.649	2.372	0.599	2.874	0.549	3.670
5	0.672	2.090	0.624	2.453	0.576	3.004
6	0.691	1.915	0.644	2.203	0.597	2.263
7	0.705	1.797	0.661	2.035	0.615	2.377
8	0.718	1.711	0.676	1.916	0.631	2.204
9	0.730	1.645	0.688	1.826	0.644	2.076
10	0.739	1.593	0.699	1.755	0.656	1.977
11	0.748	1.551	0.708	1.698	0.667	1.899
12	0.755	1.515	0.717	1.651	0.676	1.833
13	0.762	1.485	0.725	1.611	0.685	1.779
14	0.769	1.460	0.732	1.577	0.693	1.733
15	0.775	1.437	0.739	1.548	0.700	1.693
16	0.780	1.417	0.743	1.527	0.707	1.659
17	0.785	1.400	0.749	1.499	0.713	1.629
18	0.789	1.384	0.756	1.479	0.719	1.602
19	0.794	1.370	0.760	1.460	0.724	1.578
20	0.798	1.357	0.765	1.445	0.730	1.556
21	0.802	1.346	0.769	1.429	0.735	1.536
22	0.805	1.335	0.773	1.415	0.739	1.519
23	0.809	1.325	0.777	1.403	0.743	1.502
24	0.812	1.317	0.781	1.391	0.748	1.487
25	0.815	1.308	0.784	1.380	0.751	1.473
26	0.817	1.300	0.788	1.370	0.755	1.460
27	0.820	1.293	0.791	1.361	0.758	1.447
28	0.823	1.286	0.794	1.352	0.762	1.437
29	0.826	1.280	0.796	1.344	0.765	1.427
30			0.799	1.337		
40			0.821	1.279		
50			0.837	1.243		
60			0.849	1.217		
70			0.858	1.198		
80			0.866	1.183		
90			0.873	1.171		
100			0.878	1.161		

BIBLIOGRAPHY

- a. "Sample Sizes Which Give Constant Protection for Varying Lot Size", by E. L. Crow and M. W. Maxfield. NOTS Technical Memorandum RMS-5, August 1949.
- b. "Tables of Confidence Limits for Binomial Distribution", BuOrd NAVORD Report 470, 23 August 1948.
- c. "Industrial Experimentation", by Brownlee. Chemical Publishing Co., 1949. (An excellent handbook type reference.)
- d. "Introduction to Mathematical Statistics", by Hoel. Wiley and Sons, Inc., 1947. (Good theoretical discussion.)
- e. "Statistical Methods", by Snedecor. Iowa State College Press, 1946. (Sometimes referred to by the Bureau of Ordnance.)
- f. "Statistical Design and Analysis of Experiments for Development Research", by Villars. Wm. C. Brown Co., 1951. (This book includes a complete treatise on the use of a calculating machine to assist in tabulating variances.)
- g. "Introduction to Statistical Analysis", by Dixon and Massey.¹ McGraw-Hill Book Co., 1951. (Recommended by E. L. Crow as an elementary modern reference.)
- h. "Techniques of Statistical Analysis", by Eisenhart, Hastay, and Wallis. McGraw-Hill Book Co., 1947.² (Has complete tables on tolerance limits.)
- i. "An introduction to Industrial Statistics and Quality Control", by Peach. L. B. Phillips Co., 1945.
- j. "Statistical Methods Applied to Torpedo Test Results", by R. L. Deckinger². BuOrd NAVORD Report No. 29-44, no date on photostatic copy in Technical Library. (A readable treatise.) RESTRICTED.
- k. "Experimental Designs", by Cochran and Cox. Wiley and Sons, Inc., 1950. (Provides a discussion of techniques for determining the number of tests required per series to establish a difference in treatments.)

¹In the first edition, first printing, the headings on Tables 9a. and 9b. on pages 320 and 321 are interchanged.

²It is the opinion of E. L. Crow that the method used in these references for computing confidence limits for the percentage defective in a lot when no defectives are observed in the sample is not valid, in the graphs for 80 per cent and 90 per cent confidence.

1. "Statistics Manual", NAVORD in preparation to be published in 1953. (This manuscript is specifically designed for research workers and has been prepared in such a manner that it is not necessary to read material preceding that part which is of particular interest for a given experiment.)

RM/sh

Copy to:

12

15

1510

17

18

30

40

50

P80

35 (2)

3501

3502

3511

3512

3513

3521

3522

3523 (50)

3524

3525

Art Breslow, Code 4016 (12)

TEXT LOCATED AND PROVIDED

BY

BALLISTIC INJURY ARCHIVES (UK) 1977

AS A SERVICE TO THE RESEARCH COMMUNITY

Dr. Leslie D. Payne

LDPayneBIA@gmail.com